核等值：一种观察分数等值体系*

王少杰　张敏强　李拓宇　梁正妍

（华南师范大学心理学院，广州 510631）

摘　要　核等值流程包括：预平滑、估计分数概率、连续化、等值、评估等值结果。该方法兼具线性等值与等百分位等值的优点，各环节扩展性与包容性较强；采用平滑与连续化处理，可降低等值随机误差；等值差异标准误等其所特有的概念为结果评估提供可靠的工具。连续化与带宽选择方法等因素均可影响其表现；基于核等值的新方法为等值发展提供了新颖的视角。未来可关注核等值体系的扩充与完善、流程的更新、等值方法的结合和比较等方向。

关键词　核等值；连续化；带宽选择；等值新方法

分类号　**B841**

1 引言

2014 年 9 月 3 日，国务院印发的《关于深化考试招生制度改革的实施意见》指出，要“完善高中学业水平考试”，“创造条件为有需要的学生提供同一科目参加两次考试的机会”，“外语科目提供两次考试机会”。多次考试成绩间的可比性逐渐成为社会关注的热点。另一方面，教育与心理测验理论的发展与应用，使评估能力水平、兴趣爱好、职业倾向等心理特质更为方便与快捷。不同测验形式间的分数相互转换，也成为学者们研究的重点。不论是教育考试招生制度改革的社会热点，还是教育与心理测量理论的研究重点，不约而同，都在指向同样的问题——测验分数的可比性。

测验等值（Test Equating），便是解决上述问题的常用方法。它是指调整不同测验形式上的分数，使其能够相互替代的统计过程（Kolen & Brennan, 2014）。具体而言，等值是对测量同一心理特质的不同测验分数或试题参数，通过一定的数学模型，转换成同一单位系统中的量数，以利于相互比较的方法（张敏强，胡晖，1988）。常用等值方法主要包括基于经典测量理论（Classical Test Theory, CTT）的方法与基于项目反应理论（Item Response Theory, IRT）的方法。前者主要分为平均数等值（Mean Equating, ME）、线性等值（Linear Equating, LE）、等百分位等值（Equipercentile Equating, EE），后者主要分为 IRT 真分数等值（IRT True Score Equating, IRT TSE）与 IRT 观察分数等值（IRT Observed Score Equating, IRT OSE）。近些年，

收稿日期：2019-05-12

*国家社会科学基金一般项目（BHA180141）资助。

通信作者：张敏强，E-mail: 2640726401@qq.com

随着测验等值理论的发展，以核等值（Kernel Equating; 关丹丹，景春丽，2018；Dorans & Puhan, 2017; Underhill, 2017; Wallin & Wiberg, 2019; Wiberg & González, 2016）、局部等值（Local Equating; Xin & Zhang, 2015）、纳入协变量的等值方法（Equating with Covariates; González, Barrientos, & Quintana, 2015; Kim & Lu, 2018; Lu & Guo, 2018; Sansivieri & Wiberg, 2016; Wiberg & von Davier, 2017）等为代表的一批新兴等值理论与技术，为等值提供了新的视角，促进了研究与实践的发展。

最初，核等值是在美国教育考试服务中心（Educational Testing Service, ETS）的研究报告中被首次提出，其主要目的为开发新的等值方法，充分挖掘对数线性模型（Log-linear model）拟合分数分布的潜力与优势。但当时该理论并不成熟，仅适用于随机等组设计（Equivalent Groups design, EG）和非等组锚测验设计（Non-Equivalent groups with Anchor Test design, NEAT）。其后随着 von Davier, Holland 和 Thayer（2004）出版著作 *The Kernel Method of Test Equating*（《核等值》），核等值成为涵盖单组设计（Single Group design, SG）、随机等组设计、平衡组设计（Counter-Balanced group design, CB）与非等组锚测验设计的完善等值方法。近些年，研究者们又将其整合为了观察分数等值体系（observed score equating framework），进一步扩展了应用范围与价值（von Davier, 2011a, 2011b, 2013）。得益于其较 CTT 与 IRT 等值方法的突出特点与优势，核等值方法得到了国外研究者们的广泛关注（Andersson & Wiberg, 2017; Arıkan & Gelbal, 2018; De Ayala, Smith, & Norman Dvorak, 2018; Leôncio & Wiberg, 2017; Wallin & Wiberg, 2019）。但纵观国内，仅有部分学者于十余年前发表过核等值方法的研究综述并探究了它的表现（陈俊丽，2008；罗莲，2008a，2008b），尔后便无人问津。故本文旨在通过详细介绍核等值理论与操作流程，综述相关研究成果与进展，并归纳其未来研究方向，以促进其在国内的传播、普及与应用。

2 核等值理论

核等值是一种测验等值的方法体系，它基于近似传统 EE 的方法，并将 LE 作为特例。核等值研究共包含五步：（1）预平滑（Pre-smoothing），即采用对数线性模型拟合原始观察分数分布，从而得到相关的单变量或双变量分数概率分布（univariate or bivariate score probabilities）。（2）估计分数概率（Estimation of the score probabilities），即通过设计函数（Design Function），将拟合的样本分数概率转化为总体分数概率。（3）连续化（Continuization），即通过选择合适的核函数（kernel function）与带宽（bandwidth parameter），将待等值两测验

的离散累积分布函数转化为连续累积分布函数。(4) 等值 (Equating)，即采用核等值框架下的等百分位等值函数，将两测验分数进行等值。(5) 计算等值标准误 (Standard Error of Equating，SEE) 和等值差异标准误 (Standard Error of Equating Difference，SEED)，即对等值结果进行评估。下面以 EG 为例，详细介绍核等值各流程，其他等值设计及细节请参考 von Davier 等人 (2004) 的著作。

2.1 核等值流程

2.1.1 预平滑

预平滑即采用统计模型 (主要为对数线性模型) 拟合待等值两测验的样本分数分布，并通过极大似然估计 (Maximum Likelihood Estimation, MLE) 方法获得模型参数，最后经由拟合指标确定最佳模型的统计过程。

假设有待等值测验 X 与 Y，它们的分数分布为随机变量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$，用x_j和y_k分别代表其可能分数，用r_j和s_k分别代表相应分数概率，因此有向量$\boldsymbol{r} = (r_1, \dots, r_J)^t$和$\boldsymbol{s} = (s_1, \dots, s_K)^t$；用$n_j$和$m_k$分别代表对应的人数，因此有$N = \sum_j n_j$和$M = \sum_k m_k$。J 维向量$\boldsymbol{n} = (n_1, \dots, n_J)^t$和 K 维向量$\boldsymbol{m} = (m_1, \dots, m_K)^t$相互独立，且均服从多项分布 (Multinomial Distribution)，即

$$P(\boldsymbol{n}) = \frac{N!}{n_1! \dots n_J!} \prod r_j^{n_j}, \tag{1}$$

以下关于测验 Y 的性质均可由测验 X 类比得到，不作赘述。

同时，可得用于拟合分数概率的对数线性模型一般形式为

$$log(r_j) = \alpha_r + \sum_{i=1}^{T_r} \beta_{ri} (x_j)^i, \tag{2}$$

其中，α_r为标准化常数，以保证所有r_j总和为 1；T_r表示拟合的中心矩最高阶数；β_{ri}为待估计参数。

最后可通过 MLE 方法求解似然函数

$$L_r = \sum_j n_j \, log(r_j)。 \tag{3}$$

von Davier 等人(2004)提出四个评价、选择预平滑模型的标准：(1)一致性(Consistency)，即随着样本量增大，参数估计值应收敛于总体真值；(2) 高效性 (Efficiency)，即考虑到相应样本量，估计值与总体真值间的偏差应尽可能小；(3) 概率为正数 (Positivity)，即所有测验分数对应概率值均为正数；(4) 完整性 (Integrity)，即拟合的分数分布应保持与样本分数分布的中心矩 (如平均数、方差、偏度、峰度等) 不变。常用指标有 Freeman-Tukey 残差、

似然比-卡方等（Holland & Thayer, 2000）。

在此过程中，可得到关于 $\boldsymbol{r}$ 的较大维度估计协方差矩阵$\Sigma_{\hat{\boldsymbol{r}}}$。经证明，将其矩阵分解转化为较小维度的矩阵$C_r$，可应用于后续 SEE 和 SEED 的计算，提高等值分析效率。

2.1.2 估计分数概率

在估计分数概率阶段，采用与各等值设计相对应的设计函数，将预平滑阶段拟合的分数概率，经线性或非线性变换，转化为目标总体分数概率，并以向量形式表示。

von Davier 等人（2004）指出，可根据预平滑阶段参数的假设分布和数量将等值设计分为两种类型：一种是 EG，为单变量分布；另一种包含 SG、CB 和 NEAT，为双变量分布。

在 EG 中，采用设计函数得到总体分数概率的公式为

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix} = \mathrm{DF}(\boldsymbol{r}, \boldsymbol{s}) = \begin{pmatrix} \mathbf{I}_J & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K \end{pmatrix} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix}, \tag{4}$$

其中$\mathbf{I}_J$和$\mathbf{I}_K$分别为$J \times J$和$K \times K$单位矩阵。可以发现，$\boldsymbol{r}$与$\boldsymbol{s}$在转换前后并未发生改变，这是因为 EG 的特殊性，无需对样本分数概率进行转换，便可得到相应总体分数概率。为与其他三种设计的设计函数在形式上保持一致，故作此处理。而在其他三种设计中，均需首先将双变量分布概率矩阵向量化，再将类似的矩阵与其相乘，从而得到总体分数概率。

设计函数的重要作用还体现在其雅各比矩阵（Jacobian Matrix），即设计函数关于分数概率 $\boldsymbol{r}$ 与 $\boldsymbol{s}$ 的一阶偏导矩阵$\boldsymbol{J}_{DF}$，主要用于 SEE 与 SEED 的计算。

2.1.3 连续化

将待等值两测验的离散累积分布函数转化为连续累积分布函数，并将其应用于后续等值过程中，可降低因样本数据表现不稳定、不规则而导致的等值误差，这便是连续化的基本思想和操作。

经证明，将离散变量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$，与连续随机变量 $\boldsymbol{V}$ 加和，并进行一定转换，可使调整后的$\boldsymbol{X}(h_X)$与$\boldsymbol{Y}(h_Y)$连续，且中心矩在转换前后保持不变，此即核函数连续化的基本思想。常用的连续核为高斯核（Gaussian kernel）。转换后$\boldsymbol{X}(h_X)$为

$$\boldsymbol{X}(h_X) = a_X(\boldsymbol{X} + h_X\boldsymbol{V}) + (1 - a_X)\mu_X, \tag{5}$$

其中$a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2}$，$h_X$为任意正数（带宽），$\mu_X$与$\sigma_X^2$为 $\boldsymbol{X}$ 在等值总体上的平均数与方差。

接下来，为进行下一步等值操作，需得到$\boldsymbol{X}(h_X)$的累积分布函数

$$F_{h_X}(x) = \sum_j r_j \Phi\left(R_{jX}(x)\right), \tag{6}$$

其中$\Phi(z)$为标准正态分布的累积分布函数，$R_{jX}(x)=\frac{x-a_X x_j-(1-a_X)\mu_X}{a_X h_X}$。

可见，带宽h_X确定了$F_{h_X}(x)$的平滑程度。最常用的h_X选择方法为惩罚法（Penalty Method），其函数为

$$PEN(h_X)=\sum_j\left(r_j-f_{h_X}(x_j)\right)^2+K\cdot\sum_j A_j(1-B_j)\,, \tag{7}$$

其中$f_{h_X}(x)$为$\boldsymbol{X}(h_X)$的概率密度函数，且$f_{h_X}(x)=\sum_j r_j\phi\left(R_{jX}(x)\right)\frac{1}{a_X h_X}$，$\phi(\cdot)$为标准正态分布概率密度函数；$K$ 为常数，当在x_j稍偏左的位置$f'_{h_X}(x)<0$时，$A_j=1$，当在x_j稍偏右的位置$f'_{h_X}(x)>0$时，$B_j=0$。

该函数等号右边第一部分的逻辑是选择使$\boldsymbol{X}(h_X)$的概率密度函数与估计的总体r_j差异最小，且尽可能拟合原始分布中的“齿状（teeth）”与“跳跃（gaps）”形态的带宽；第二部分的逻辑为惩罚使概率密度函数呈现“U”形分布的带宽，以确保连续化后的分数分布平滑，最后选择最小惩罚函数值对应的h_X作为带宽。

2.1.4 等值

借助上述结果，依据 EE 的基本思想，可得到核等值函数的一般表达式为

$$e_Y(x)=e_Y(x;\boldsymbol{r},\boldsymbol{s})=G_{h_Y}^{-1}\left(F_{h_X}(x;\boldsymbol{r});\boldsymbol{s}\right)=G_{h_Y}^{-1}\left(F_{h_X}(x)\right)\,, \tag{8}$$

其中$F_{h_X}(x;\boldsymbol{r})$为连续化后$\boldsymbol{X}(h_X)$的累积分布函数，$G_{h_Y}^{-1}(\cdot)$为$\boldsymbol{Y}(h_Y)$的累积分布函数的反函数。对于分数$x_j$，可首先找到其在$\boldsymbol{X}(h_X)$总体上的百分等级，进而求得该 百分等级在$\boldsymbol{Y}(h_Y)$总体上对应的百分位数，即等值分数$e_Y(x_j)$。

经 von Davier 等人（2004）证明，LE 与 EE 间相差一个形状差异函数 R(x)（shape difference function）。可选择较大带宽（通常为$h_X>10\sigma_X$，$h_Y>10\sigma_Y$），使核等值结果近似于 CTT 等值中的 LE 结果。此即核等值将 LE 作为其特例的理论依据。

2.1.5 计算 SEE 与 SEED

SEE 为等值随机误差，主要来源于抽样方法。假设等值基于目标总体而非样本，便不存在 SEE。可采用 δ 方法计算 SEE，其基本公式为

$$SEE_Y(x)=\sigma_Y(x)=\sqrt{Var\left(e_Y(x)\right)}。 \tag{9}$$

特别地，核等值 SEE 的计算通常采用如下方式

$$SEE_Y(x)=\|\boldsymbol{J}_{e_Y}\boldsymbol{J}_{DF}\boldsymbol{C}\|\,, \tag{10}$$

其中，$\boldsymbol{J}_{e_Y}$为核等值函数关于 $\boldsymbol{r}$ 与 $\boldsymbol{s}$ 的雅各比矩阵；$\boldsymbol{J}_{DF}$为设计函数关于 $\boldsymbol{r}$ 与 $\boldsymbol{s}$ 的雅各比矩阵，

$\| \boldsymbol{v} \| = \sqrt{\sum_j v_j^2}$，即向量$\boldsymbol{v}$的欧几里得范数；$\boldsymbol{C}$为估计分数概率阶段得到的$C_r$矩阵。

对于相同等值数据，采用不同等值方法所得结果间差异的标准差，即为 SEED，其主要用于核等值函数间的比较，作为衡量其差异程度的指标。只有当等值函数间差值在$[-2SEED, 2SEED]$外时，才可认为其差异显著。计算公式为

$$SEED_Y(x) = \sqrt{Var\left(e_1(x) - e_2(x)\right)} = \| \boldsymbol{J}_{e_1} \boldsymbol{J}_{DF1} \boldsymbol{C} - \boldsymbol{J}_{e_2} \boldsymbol{J}_{DF2} \boldsymbol{C} \| , \tag{11}$$

各参数含义同 SEE。

2.2 核等值特点

核等值主要有六个特点：（1）它将 CTT 等值中最常用的 LE 与 EE，作为特例，纳入统一框架，扬其长避其短；（2）理论体系完善，从等值设计到等值评价，均在设计函数等一系列核等值所特有且相互联系的框架中完成；同时也可对各环节单独分析（模块化），便于等值的评估与改善；（3）可调整各参数，形成不同等值方法，极具包容性与扩展性；（4）开创性地提出 SEED 的概念，作为等值结果间差异的比较基准；（5）采用预平滑和连续化，可显著降低因样本量过少造成的等值随机误差，同时也适于大样本等值；（6）并未限定待等值测验间极端分数对应等值（EE 与 IRT OSE 均有此不足），而是根据核函数将等值分数范围扩展。

2.3 核等值评价指标

评价核等值的指标主要包括 SEE、SEED 与 PRE。SEE 与 SEED 前已介绍。PRE 旨在度量等值后的$e_Y(\boldsymbol{X})$与$\boldsymbol{Y}$的分布差异，从而判断其对“完整性”的满足程度，公式如下

$$PRE(p) = 100 \times \frac{\mu_p\left(e_Y(\boldsymbol{X})\right) - \mu_p(\boldsymbol{Y})}{\mu_p(\boldsymbol{Y})}, \tag{12}$$

其中$\mu_p(\boldsymbol{Y}) = \sum_k (y_k)^p s_k$，$\mu_p\left(e_Y(\boldsymbol{X})\right) = \sum_k \left(e_Y(x_j)\right)^p r_j$。

其他在等值领域通用的评价指标也适用于核等值。例如，bias、DTM、MAD、RMSD（Kolen & Brennan, 2014）与 RMSE（Wallin & Wiberg, 2019）等。

3 研究进展

3.1 观察分数等值体系

基于核等值研究的五个流程，von Davier（2011a, 2011b, 2013）提出了观察分数等值体系。她认为，选择不同的对数线性模型、等值设计和带宽，可分别改变公式（10）中的$\boldsymbol{C}$矩

阵、$\boldsymbol{J}_{DF}$和$\boldsymbol{J}_{e_Y}$，进而改变等值结果，使核等值从单一的方法扩展为体系。例如，von Davier 等人（2006）提出，可通过调整带宽，使核等值结果近似于部分 CTT 等值结果，详见表 1。

可见，由核等值发展而来的观察分数等值体系具有灵活、可拓展等优势，为其理论的扩充及与其他等值方法的结合提供便利。

表 1 常用 CTT 等值与核等值方法对应表

等值设计	CTT 等值	核等值
EG	等百分位等值	核等值（最优带宽）
	线性等值	核等值（较大带宽，$h_X > 10\sigma_X$，下同）
NEAT	等百分位链等值	核链等值（最优带宽）
	等百分位后分层等值	核后分层等值（最优带宽）
	线性链等值	核链等值（较大带宽）
	Tucker 等值	核后分层等值（较大带宽，特定条件下）
	Levine 观察分数等值	-

3.2 连续化方法

3.2.1 Epanechnikov 核

Cid 和 von Davier（2015），González 和 von Davier（2016）将 Epanechnikov 核引入核等值研究，它通过赋予近分数点区域较大权重，远分数点区域较小权重，从而在处理有界变量（bounded variables）时更具优势。具体而言，Epanechnikov 核的概率密度函数和累积分布函数分别为

$$f(v) = \frac{3}{4}(1 - v^2) \qquad |v| \le 1, \tag{13}$$

$$F(v) = \begin{cases} 0 & v < 1 \\ \dfrac{3v - v^3 + 2}{4} & -1 \le v \le 1。\\ 1 & v > 1 \end{cases} \tag{14}$$

进而，可得连续化累积分布函数

$$F_{h_X}(x) = \sum_{\substack{j: \\ -1 \le R_{jX} \le 1}} \frac{r_j\left(3R_{jX}(x) - R_{jX}^3(x) + 2\right)}{4} + \sum_{\substack{j: \\ R_{jX} > 1}} r_j, \tag{15}$$

其余各参数和操作均与高斯核连续化相同，不作赘述。

3.2.2 自适应核

同样是 Cid 和 von Davier（2015），González 和 von Davier（2016），将自适应核（Adaptive kernel）引入核等值研究。与高斯核不同，自适应核可依概率密度调整带宽。例如，在低密度值（极端分数）处，选择较大带宽，以使分数分布更为平滑，减小等值误差。主要分为三步：

（1）根据 2.1.3 连续化的思路，求得初步$f_{h_X}(x_j)$。

（2）求取各分数点处带宽权重系数，$\lambda_j = \left(\frac{f_{h_X}(x_j)}{g}\right)^{-\alpha}$。其中，$g$为所有分数点$f_{h_X}(x_j)$的几何平均数；$\alpha$为稳定系数，且$-1 \leq \alpha \leq 1$，一般取$\alpha = 0.5$。

（3）可得连续化后的累积分布函数为

$$F_{h_{jX}}(x) = \sum_j r_j \Phi\left(\frac{x - a_{jX}x_j - (1 - a_{jX})\mu_X}{a_{jX}h_{jX}}\right), \tag{16}$$

其中，$a_{jX}^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_{jX}^2}$，$h_{jX} = \lambda_j h_X$。

Cid 和 von Davier（2015）模拟生成了包含不同形态（对称、正偏、负偏、两种轻微负偏）和极端分数占比（百分位数 P2.5 或 P97.5 的占比为 4%和 8%，最小或最大分数的占比为 4%[†]）的作答数据，以比较传统高斯核、Epanechnikov 核和自适应核在不同分数分布和极端分数情况下连续化的结果。他们发现，Epanechnikov 核在两种轻微负偏分布尾端的平滑效果较好。当包含 8%极端分数时，自适应核与高斯核表现相似；当最小或最大分数占比为 4%时，在极端分数处，自适应核平滑效果较另外两种方法好。而传统高斯核对尖状分布（spike）的拟合效果更好。可见，连续化方法在极端分数处的表现受分数分布形态影响。

3.2.3 其他方法

Logistic 核和均匀核（Uniform kernel）是两种较为传统的连续化方法，它们各自采用 Logistic 分布和均匀分布的核函数，对等值样本分数进行连续化处理。

Logistic 核的累积分布函数和概率密度函数分别为

$$H(v) = \frac{1}{1 + exp\,(-v/s)}, \tag{17}$$

$$h(v) = \frac{exp\,(-v/s)}{s(1 + exp(-v/s))^2}, \tag{18}$$

其中，s 为量尺参数（scale parameter）。经 Logistic 核连续化后$\boldsymbol{X}(h_X)$的累积分布函数与概率

[†]在该研究中，设定“百分位数 P2.5 或 P97.5”的意义为，依据分数分布形态（对称分布除外）取其一。比如，模拟正偏态分布，极端分数便为 P97.5。最小或最大分数占比同理。

密度函数分别为

$$F_{h_X}(x) = \sum_j r_j H\left(R_{jX}(x)\right), \tag{19}$$

$$f_{h_X}(x) = \sum_j r_j h\left(R_{jX}(x)\right)\frac{1}{a_X h_X}。 \tag{20}$$

同样的思路，可得均匀核的累积分布函数和概率密度函数、经均匀核连续化后$\boldsymbol{X}(h_X)$的累积分布函数与概率密度函数分别为

$$H(v) = \begin{cases} 0 & v < -b \\ \dfrac{v+b}{2b} & -b \le v < b, \\ 1 & v \ge b \end{cases} \tag{21}$$

$$h(v) = \begin{cases} \dfrac{1}{2b} & -b < v < b \\ 0 & \text{其他} \end{cases}, \tag{22}$$

$$F_{h_X}(x) = \sum_{\substack{j: \\ R_{jX}(x) \ge b}} r_j + \sum_{\substack{j: \\ -b \le R_{jX}(x) \le b}} \left(r_j \cdot \frac{R_{jX}(x)+b}{2b}\right), \tag{23}$$

$$f_{h_X}(x) = \sum_{\substack{j: \\ -b \le R_{jX}(x) \le b}} \frac{r_j}{2b}。 \tag{24}$$

Lee 和 von Davier（2008, 2011）将 EG 中采用高斯核平滑的核等值作为参照基准，比较了 Logistic 核与连续均匀核的表现。结果证明高斯核对原始数据高阶中心矩具有良好返真性，核函数的尾部形态对连续化方法影响较大。例如，高斯核与 Logistic 核的核函数在整个分数区间均为正数，而连续均匀核并非如此。除此之外，采用不同连续核的等值结果间差异不大。

3.3 带宽选择方法

3.3.1 重复平滑法

Häggström 和 Wiberg（2014）将重复平滑法（Double Smoothing method）应用于带宽选择中，并将其与惩罚法比较，最后发现二者结果相似。顾名思义，重复平滑即进行两次平滑处理，以减小数据离散导致的等值误差。重复平滑法分为三步：

（1）以原始数据最小分数点的一半为单位进行连续化处理，得到$\boldsymbol{X}(h_X)$的概率密度函数

$$f_{g_X}(x) = \sum_{j=1}^{2J-1} r_j \phi\left(R_{jgX}(x)\right)\frac{1}{a_{gX} h_{gX}}, \tag{25}$$

其中，$R_{jgX}(x) = \frac{x - a_{gX}x_j - (1-a_{gX})\mu_X}{a_{gX}h_{gX}}$，$a_{gX}^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_{gX}^2}$。

（2）利用$f_{g_X}(x)$，再次计算$\boldsymbol{X}(h_X)$的概率密度函数

$$f_{h_X}(x) = \sum_{j=1}^{J} f_{g_X}(x_j)\phi\left(R_{jX}(x)\right)\frac{1}{a_X h_X}。 \tag{26}$$

（3）计算使得重复平滑函数取得最小值的带宽，该函数为

$$DS(h_X) = \sum_{l=1}^{2J-1}\left(\hat{r}_l - f_{h_X}(x_l)\right)^2， \tag{27}$$

其中，$\hat{r}_l = \begin{cases} r_{\frac{l+1}{2}}, & l\text{为奇数} \\ f_{h_X}(x_l), & l\text{为偶数} \end{cases}$。

3.3.2 交叉验证法

Liang 和 von Davier（2014）提出的交叉验证法（Cross-Validation method），先将样本分成两部分，然后将基于样本所得的泊松似然函数最大化，以获得对应的最优带宽。具体分为四步：

（1）将数据分成两个随机子样本。

（2）采用样本 1 计算$F_{h_X}(x)$和$f_{h_X}(x)$。由于h_X未知，取一定范围的值作为初始值（例如在$[0.01, 5]$每隔 0.01 取h_X）。因此，对于每个h_X，在任一分数点处均可找到对应值。

（3）假设频率服从泊松分布，交叉验证过程可通过该分布体现，即

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}， \tag{28}$$

其中，λ为步骤（2）中对于给定h_X，特定分数的概率密度值；k为样本 2 中该分数的频率。

（4）将各分数的概率值相乘得到似然函数，并取自然对数。该函数最大值所对应的h即为一个最优带宽。将此过程重复 1000 次并求取中位数得到最终带宽。

他们将采用交叉验证法和惩罚法的核等值与两种 EE 结果对比，发现在 bias、SEE 和 PRE 角度，采用交叉验证法的核等值优于后两种等值方法。

Wallin, Häggström 和 Wiberg（2017）认为，交叉验证法需重复计算，运算效率低，为此提出了删一交叉验证法（Leave-One-Out Cross-Validation method）。在计算$f_{h_X}(x_j)$时，该方法将x_j及其对应频率删除，从而减小模型过拟合问题，提高了运算效率。研究发现，从等值分数角度，带宽选择方法的表现彼此相似；但在高分段，不同方法所得等值结果间存在较大差异。

3.3.3 Silverman 经验准则

Silverman 经验准则通过使渐近平均积分平方误差（asymptotic mean integrated squared

error）最小化，求取对应带宽（Andersson & von Davier, 2014）。当分数为正态分布且采用高斯核平滑时，可得到 Silverman 经验准则为

$$h_X \approx 1.06\sigma_X n_X^{-\frac{1}{5}}。 \tag{29}$$

他们认为，可用 0.9 代替 1.06，以减小异常值的影响，避免数据过度平滑。同时，考虑到a_X可影响最优带宽，调整的 Silverman 经验准则为

$$h_X = \frac{9\sigma_X}{\sqrt{100n_X^{\frac{2}{5}} - 81}}。 \tag{30}$$

研究表明，当分数分布较为平滑时，惩罚法第一部分表现优异；反之，调整的 Silverman 经验准则较好。（调整的）Silverman 经验准则直接采用公式，计算得到带宽，简单、直接；但当正态分布假设不满足时，该方法可能会带来较大误差。

3.4 基于核等值的新方法

3.4.1 纳入协变量的核等值

当样本组间存在明显的能力差异时，在等值中，一般使用锚测验调整两次测验间的难度差异。但锚题高曝光率，又使测验保密成为难点。为此，有学者提出，在非等组条件下，利用人口学信息（例如年龄、性别、教育背景等）调整组间能力差异，从而构造出伪等组（Pseudo-Equivalent Groups）考生进行等值（Haberman, 2015）。研究发现，该方法的表现与其他等值方法不相上下，甚至可能更胜一筹（González et al.,2015; Kim & Lu, 2018; Longford, 2015; Lu & Guo, 2018; Sansivieri & Wiberg, 2016）。其流程大致分为以下三步：

（1）构建目标背景变量分布。假设Z_{iX}和Z_{jY}分别为作答测验 X 与测验 Y 的第 i 位和 j 位考生的背景变量。那么，可得测验 Y 背景变量平均数为

$$\bar{Z} = \sum_{j=1}^{N_Y} z_{jY} \Big/ N_Y\,, \tag{31}$$

其中，N_Y为参与测验 Y 的考生人数。采用最小区分信息法（minimum discrimination information method; Haberman, 1984）调整权重w_{iX}，使得

$$\sum_{i=1}^{N_X} w_{iX} z_{iX} \Big/ N_X = \bar{Z}, \tag{32}$$

其中，$w_{iX} > 0$，且$\sum_{i=1}^{N_X} w_{iX} = 1$。

（2）计算伪等组的分数分布。依据求得的权重w_{iX}，获得考生作答测验 X 的伪等组分数分布。

（3）采用 EG 下的等值方法对测验 X 伪等组分数与测验 Y 分数进行等值。

Wiberg 和 Bränberg（2015）将该方法引入核等值，用背景信息作为协变量，替代 NEAT 中的锚测验，其余操作流程与核等值基本相同。研究发现，该方法可与 EG 等值相媲美；同时使用锚测验和背景信息，可获得更为准确的等值结果。不过，当背景变量的水平组合过多时，每个组合上的考生人数会急剧减少，影响等值结果。为此，Wallin 和 Wiberg（2016, 2019）提出，用 Logistic 回归函数替代原来将背景变量简单加和的方法，并得到用以匹配考生的倾向分数（Propensity Score）。结果表明该方法与采用锚题等值的结果相似，较同等条件下 EG 等值的结果好，在一定程度上解决了变量水平组合过多带来的问题。

3.4.2 局部观察分数核等值

局部等值依据考生能力差异，分别构建一族等值函数，从而做到等值因“能力”而异，得出更为精确的结果（van der Linden, 2010, 2013）。该方法的提出源于 Lord（1980）对等值公平性的定义，即分数分布在等值转换前后保持不变。传统等值方法将单一转换关系应用于整个等值群体的思路并不完全合适，主要因为等值关系依赖于特定总体且存在偏差（Wiberg, 2016a）。为此，依据考生能力不同，局部等值构建了一系列转换关系。假设考生能力已知，可使用测验 X 与 Y 的条件分布计算相应的局部等值，即

$$\varphi^*(y;\theta) = F_{X|\theta}^{-1}\left(F_{Y|\theta}^{-1}(\mathrm{y})\right)。 \quad (33)$$

需要注意，这里的θ代指特定能力考生，可以为锚测验分数、通过 IRT 求得的能力值θ等。

众多研究表明，局部等值的精度与 IRT 等值方法相近，且均优于 CTT 等值方法（van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011; Xin & Zhang, 2015）。受此启发，Wiberg, van der Linden 和 von Davier（2014）将局部等值与核等值相结合，提出三种局部观察分数核等值方法，并将其与局部等值和核等值对比。结果发现，局部 IRT 观察分数核等值方法的 bias、PRE 和 SEE 均较小；局部核等值方法的 bias 较小，准确性受锚测验长度影响不大，结果较为稳定，但 SEE 较大。综合考量，他们认为，局部 IRT 观察分核等值可替代 IRT 观察分数核等值。Wiberg（2016a）提出了线性 IRTOSE，该方法主要思路是用 IRT 模型拟合作答数据，然后利用作答反应概率求得 CTT 线性等值中的总体参数进行等值。她比较了 IRTOSE、线性 IRTOSE、局部线性 IRTOSE 与局部线性 IRTOSE 核等值四种等值方法，发现与 IRTOSE、线性 IRTOSE 方法相比，两种局部等值方法在 MSD、MAD、RMSD 等指标上的表现较为优异。

3.4.3 IRT 观察分数核等值

IRT 观察分数核等值采用 IRT 模型拟合测验原始数据，以获得两测验相应的得分概率。其余步骤与核等值的第三步至第五步基本相同。Andersson, Bränberg 和 Wiberg（2013）首先提出该方法，并详细介绍了其在 NEAT 中的操作流程。随后，Andersson（2016）推导出多级计分 IRT 观察分数核等值的渐近标准误（asymptotic standard error），其主要用于采用分析方法计算 SEE 的过程。经研究证明，在不同样本量、能力分布、锚测验长度条件下，该标准误的估计均较为准确，为相关研究中等值表现的评估提供了可靠指标。Wiberg（2016b），Andersson 和 Wiberg（2017），Sansivieri, Wiberg 和 Matteucci（2017）均从不同角度开展过相关研究，论证了 IRT 观察分数核等值的良好表现。

3.5 核等值与常用等值方法的比较研究

关于核等值与常用 CTT、IRT 等值方法的比较研究，学界并未达成共识；但受益于其独特的优势，核等值方法获得学者们较多关注与青睐。现综述如下：

有研究表明，核等值的表现优于 CTT 与 IRT 等值方法。例如，von Davier 等人（2004）在核等值框架下提出了处理 CB 数据的新方法，即两个独立单组设计。该方法可通过调整两个 SG 等值对最终等值的合成权重，以获得不同等值结果，并且可将传统视 CB 为两个 SG 或 EG 的处理方式，作为特例来处理。当等值数据存在顺序效应时，他们发现该方法的 SEE 更小，从而验证了核等值体系的优异性。von Davier 和 Chen（2013）指出，目前共有三种处理 NEAT 数据的方法。第一种是将锚测验分数作为条件变量（conditioning variable），主要分为 Tucker 等值与后分层等值方法（Post-Stratification Equating，PSE）；第二种是将锚测验分数作为待等值两测验间链接的桥梁，主要为链等值方法（Chain Equating, CE）；第三种是将锚测验分数与 CTT 相结合，主要为 Levine OSE 方法。根据第三种思路，在核等值框架下，他们提出了基于锚测验真分数的混合 Levine EE 与 PSE。研究表明，该方法的等值表现优于 CTT PSE 与 CE。在等值研究中，各方法均有适用的条件与范围。假如数据不满足等值前提，便可能会误导研究者，甚至得出错误的结论。基于此，为确保等值结论可靠，Arıkan 和 Gelbal（2018）首先验证了等值前提假设的满足程度，在此基础上，经由 EG 同样发现了核等值优于 CTT LE 与 EE 方法。为避免数据模型对等值方法的偏向，Norman Dvorak（2009）采用较为中立的因素分析模型（factor-analytic model），通过蒙特卡洛模拟生成考生作答反应数据，比较了核等值与测验特征曲线（Test Characteristic Curve, TCC）等值方法的表现，发现前者在中间分数与高分数处表现较好，后者在中间分数处表现较差。他认为，中间分数段通常是

重要决策（通过、合格）的依据点，所以核等值方法更优。Wedman（2017）也发现，在 SG 下，将核等值与 CTT 和 IRT 等值方法应用于瑞典学业测试的等值中，前者的表现优于后者。De Ayala 等人（2018）在 NEAT 下，比较了核等值与 TCC 的表现。结果表明，两种方法等值结果准确性均较高，但在参数真分数量尺上，核等值的表现优于后者，且具有较高的灵活性。

但也有研究表明，核等值与其他等值方法不分伯仲。例如，von Davier 等人（2006）在 NEAT 内锚与外锚设计下构造伪测验，并将 EG 等值结果作为参照标准（criterion equating），发现核等值与 LE、EE 方法的误差相当，且前者更接近预先设定的等值标准。Liu 和 Low（2007）的研究也得出了相似的结论。另一方面，有学者尝试采用核等值框架开发新的等值方法，同时比较其与常用等值方法的表现，从而推测核等值体系的普遍特性，不失为别具一格而又颇有价值的着眼点。例如，von Davier, Fournier-Zajac 和 Holland（2007）指出，与常用 LE 方法相比，Levine OSE 凭借其较小的误差，被频繁应用于等值实践中。但在 EE 领域，却一直未有与之相对应的等值方法。因此，他们在核等值框架下，通过整合 Levine OSE 与 EE，提出了混合等值函数（hybrid equating functions）方法，并将其应用于实证研究中，最后发现该方法与 CTT 等值方法的结果非常相似。Grant, Zhang 和 Damiano（2009）以及 Chen（2012）在 NEAT 下，比较了 IRT OSE 与基于核等值的 Levine OSE 方法。他们均发现，基于不同理论假设的两种等值方法所得结果十分相似，可将基于锚测验真分数的 PSE 视为线性 Levine OSE 方法。采用瑞典学业测试及巴西国家基础教育测试数据，Leôncio 和 Wiberg（2017）比较了 IRT OSE、核等值与 IRT OSE 核等值这三种等值方法，发现基于 IRT 的等值结果较稳定、准确，但核等值效率更高；如果选择合适的模型拟合考生分数分布，核等值结果会较为理想。

更多研究发现，在不同条件下，核等值与其他等值方法的表现各有优劣。例如，Choi（2009）通过模拟研究，生成不同测验长度与样本量的考生作答数据，进而比较了核等值与 EE 方法的表现差异。最后发现，在 EG 下，二者表现相当；但在 NEAT 下，只有 PSE 核等值与 CTT 等值方法结果相当。Meng（2012）通过模拟研究，操纵样本量、锚测验长度、能力水平三个自变量，进而比较了 PSE 核等值与 IRT 等值方法的表现。结果表明，随着锚测验长度和样本量的增加，等值误差逐渐降低；能力水平对等值结果影响较大；在不同分数区间，等值方法表现各异。概括而言，核等值方法稳定性较好，但不如 IRT 等值方法准确。

读者可能会质疑：样本量作为自变量都出现在了上述模拟研究中，为何其对结果的影响各异？比较后可发现，除样本量外，Choi 操纵了测验长度，而 Meng 操纵的是锚测验长度和

能力水平。故有理由推测，自变量数量与水平设置存在差异，导致自变量间产生不同类型与程度的交互作用，可能是研究结论间不尽相同，甚至相互矛盾的直接原因。

例如，同样为比较不同等值方法在极端分数处的表现，Godfrey（2007）发现在多数情况下，核等值方法较稳定、准确，但在极端分数处，其偏离参照等值较多，表现不如等百分位链等值（Chained Equipercentile Equating, CEE）与 IRT TSE；Moses, Yang 和 Wilson（2007）却证明，在中间分数段，核等值与 EE 结果相似；但是在极端分数处，核等值方法表现较好。比较它们的研究设计，可找到导致二者截然不同结论的可能原因——Godfrey 在 NEAT 中保持锚测验难度不变，模拟生成了不同测验难度、样本量、锚测验长度的数据，并将 SG 下的 EE 作为参照等值；而 Moses 等人是通过一批实证数据开展等值研究，并未操纵变量水平，故无法保证其结论的外部效度及两次研究结论的可比性。同样的道理，罗莲（2008a）发现核等值与 CTT 等值方法均有较好的表现；在小样本情况下，等值结果间的差异较小；但在大样本情况下，表现出较大差异。而陈俊丽却发现，当以 SG 下核等值结果作为参照等值时，NEAT 下的核等值方法表现最优，EE 表现最差；而当以 SG 下的 LE 结果为参照等值时，NEAT 下的 LE 方法表现最优，EE 方法依然表现最差。可见，两研究间的预设条件并不相同，当测验难度及考生水平均存在差异的情况时，罗莲将 EG 下的 EE 作为参照等值；而陈俊丽将目光转向设定不同的参照等值，以比较等值方法间的表现。

不可否认，虽然以上逻辑推论具有一定的合理性，但这些仅为推测。多个因素对等值结果的影响方式及结果，仍需更多实证与模拟研究进行验证与支撑，进而为比较核等值与其他等值方法间的表现，提供更多具有说服力的证据。

3.6 核等值的影响因素

影响核等值表现的因素主要有：（1）核等值相关变量，主要包含预平滑模型、连续化方法、带宽选择方法；（2）待等值群体间相关变量，主要包含群体间的能力表现差异、分数分布形态、样本量；（3）等值设计相关变量；（4）待等值测验相关变量，主要包含测验长度、锚测验长度等。其中，连续化方法和带宽选择方法前已涉及，详见 3.2 与 3.3 部分。

3.6.1 预平滑模型

作为核等值研究的第一步，通过拟合测验原始分数分布，得到用于后续计算的矩阵，预平滑模型的拟合与选取无疑会对等值结果产生不可忽视的影响。例如，Godfrey（2007）将 SG 下的 EE 作为参照等值，探讨了对数线性模型对核等值的影响。结果发现，除具有小样本、长测验特征的数据外，6-6-4 模型（即保持测验原始分数分布的六阶中心矩与四阶交叉

矩不变）均能够较好地拟合所有作答数据，获得准确的等值结果。Moses 和 Holland（2007）基于实测数据，构建伪测验及考生作答反应，比较了对数线性模型对核等值准确性的影响。他们发现，模型拟合准确性直接影响等值误差，但仅有 2-2-1 最简模型对等值结果有实际影响，其他模型影响不大。以上研究均一定程度表明，对数线性模型的拟合性能，均可能影响核等值准确性，但鉴于当前对数线性模型拟合指标等领域的研究愈发成熟与完善，其对等值结果的影响反而不甚明显。例如，在 NEAT 中，Kim（2014）比较了对数线性平滑方法及三次样条后平滑方法（cubic spline postsmoothing）对四种等值方法（PSE、修正的 PSE、CEE、核等值）的影响，发现虽然采用三次样条后平滑方法的核等值随机误差较小，但经由对数线性平滑后的等值系统误差及总误差均较前者小。可见，不再拘囿于对数线性模型，探索将更多数据平滑方法引入核等值，不失为一创新点。

3.6.2 等值设计

常用等值设计主要有 SG、EG、CB 与 NEAT。待等值考生间的能力差异，成为区分等值设计的关键因素；采用不同等值设计处理相同或相似的考生作答数据，结果可能并不相同。例如，Kim（2014）比较了 NEAT 内锚与外锚两种等值设计中四种等值方法（PSE、修正的 PSE、CEE、等百分位核等值）的表现差异，发现在相同条件下，外锚设计的等值误差，尤其是随机误差，较内锚设计小。为更好地处理等值子群体异质问题，Duong 和 von Davier（2008）在核等值框架下提出了“混合分布平衡设计（Mixture Distribution Counter-Balanced design）”的思路。该方法吸纳 SG 与 CB 的优势，首先用对数线性模型拟合不同子群体分数分布，再采用 CB 设计处理等值数据，调整权重，从而获得最优等值结果。他们采用 2PL 模型模拟作答数据，将 IRT 多群组校准真分数等值作为参照基准，比较了核等值中五种不同的处理方法。最后发现，当权重系数与样本量对应成比例时，混合分布平衡设计比其他等值方法误差小，且当子群体间能力差异较大时这种优势尤为明显。

但也有学者提出了相反意见。例如，Jiang, von Davier 和 Chen（2012）采用模拟的外锚测验数据及实测内锚测验数据，探究了等值总体合成权重对 PSE 和 CE 的影响。他们发现，各等值方法表现相似；具体而言，在 CE 中，测验 X 与锚测验 A 链接（Linking）结果的 PRE 大于锚测验 A 与测验 Y 链接结果的 PRE。这是因为将长测验（测验 X）分数分布转化为短测验（锚测验 A）分数分布更为困难。总之，一方面，可思考如何更为科学地比较常用等值设计下核等值间的表现；另一方面，可探索新的等值设计思路，为核等值，甚至其他常用等值方法提供更好的切入点。

3.6.3 待等值总体间能力差异

一般认为，采用不同等值方法处理等值数据，其结果的相似性与待等值总体间能力差异有关：当能力差异较大时，等值结果不尽相同；相反，等值结果基本一致（Dorans, Liu, & Hammond, 2008; Holland, von Davier, Sinharay, & Han, 2006; Sinharay & Holland, 2010; Wang, Brennan, & Kolen, 2008）。这是因为，在等值研究中，存在两种测验分数差异来源，它们分别由测验本身难度或考生总体间能力水平不同所致。若要保证等值准确性，就必须将不同能力水平带来的分数差异从测验分数差异中分离。对于 SG、EG，当考生总体间能力差异较大时，本身就已违反"能力水平相同"的前提假设；对于 NEAT，各种处理方式（例如 Levine 等值、Tucker 等值）的准确性也都间接地依赖于总体间分数（能力）水平分布相似的假设。从而，它们均在一定程度上受到能力差异的影响。在实证研究中，Liu 和 Low（2007, 2008）基于相同和不同年份的两套 SAT 口语测试数据分别构造出能力差异小（similar population）与能力差异大（distant population）的待等值总体，进而比较了核等值与 CTT 等值方法的表现。结果发现，当总体间的能力差异较小时，各等值结果相似；相反，当总体间的能力差异较大时，其结果并不一致。采用相似研究设计，Duong 和 von Davier（2008）也发现，当子群体间存在较大能力差异时，采用混合分布平衡设计所得的等值结果较传统 CB 处理方式所得结果更为稳定。Kim（2014）采用 IRT 模型模拟作答数据，发现当组间平均能力差值为 0.05、0.2 或 0.5 时，采用预平滑处理优于采用后平滑处理的等值表现；当差值为-0.2 时，结果则相反。他推测，这可能与模型对能力差异的假定不同有关。

3.6.4 样本量

样本量可从侧面刻画出等值结果的稳定性，即随机误差大小。当其他条件不变时，增大样本量可减小随机误差。例如，Godfrey（2007）探讨了样本量、对数线性模型、测验难度及锚测验长度对核等值结果的影响。他将 SG 下的 EE 作为参照等值，以探究包括核等值、CTT 等值与 IRT 等值方法间的表现差异，最后发现，当样本量较小时，不同等值方法间存在明显差异；但随着样本量的增大，等值误差逐渐减小，各等值方法所得结果趋于一致。Moses 和 Holland（2007）基于真实测试数据，构建伪测验及考生作答反应，比较了对数线性模型拟合准确性与样本量对核等值准确性的影响。他们发现，样本量越大，标准误的估计准确性越高。Kim（2014）、Liang 和 von Davier（2014）也得出了类似结论。这是因为，等值误差包含系统误差与随机误差，系统误差主要来源于估计准确性、统计假设、等值设计与组间差异，其随样本量变动较小；而主要来源于抽样代表性的随机误差则不然，它会随样本量增大而减小；

进而，如果等值基于目标总体数据而非样本数据，便不存在随机误差。当然，这种完全理想化的情况在实际中不可能实现。故而，增加样本量对等值有积极影响。

3.6.5 测验长度

不计其他因素，测验题目数量越多，信度也就越高，从而为等值提供了有利前提；但题目数量增多导致分数区间增大，在同等情况下，又使得分配到各分数点上的考生数量减少，进而可能会增大等值误差（Wang et al., 2008），似乎很难衡量二者孰大孰小。但实际情况是，随测验题目数增多，高信度对等值的影响增量较小，而各分数频率却会显著减小，从而使后者发挥了举足轻重的作用。例如，Norman Dvorak（2009）开展模拟研究比较了核等值与 TCC 的表现，总体来看，随着测验题目数量从 25 道增加到 75 道，测验信度不断增加，而均方根差异等误差指标值也在随之增大。

然而，在 NEAT 中，锚测验题目增多却可能会减小等值误差。这主要是因为，锚测验是区分考生能力差异与试卷难度差异的关键因素，适当增加长度，可提高其对不同变异源所导致的等值差异的区分能力，进而提高等值准确性。例如，Kim（2014）的模拟研究发现，增加锚题比例可有效减小等值误差，并以系统误差变化最为明显。Andersson（2016）也发现，相较短锚测验，CE 核等值与 PSE 核等值在长锚测验条件下均具有更为稳定且较小的标准误；由于 CE 对锚测验长度依赖较少，也使得其表现优于 PSE，进而从另一角度也验证了上述假设。但锚测验长度对等值准确性的影响也存在一定程度的边际递减效应，即随着锚题数量增多，等值准确性的改善空间逐渐减小，甚至可能会停滞。例如，De Ayala 等人（2018）探究了测验长度（25 题、50 题、100 题）和锚题占比（10%、20%、30%）对核等值和 TCC 等值准确性的影响，发现除了考生能力估计外，等值结果几乎不受二者影响。他们推测，这主要是因为在其研究中，整体来看，测验信度改观较小，因而锚测验长度的变化对等值结果影响甚微。

3.6.6 分数分布特征

为减小因分数分布不规则带来的误差，在预平滑阶段，核等值采用对数线性模型拟合样本分数分布；在连续化阶段，采用核平滑方法使离散的累积分布函数连续化。故可认为，样本分数分布特征可能会影响核等值的表现，即，其在分数分布频率较小处的表现不如在分数分布频率较大处的表现。例如，Underhill（2017）将等值总体分数分布纳入研究范畴，采用模拟方法，操纵样本量及核等值的预平滑程度，探究了在 EG 设计、样本具有不同程度的非正态分布情况下，核等值的稳健性。结果表明，在分数分布频率较高处，核等值较稳定、准

确；相反，在频率较低（例如极端分数）处，核等值波动较大。Mao（2006）的研究表明，分数平滑程度影响 SEE 估计。具体表现为，当预平滑阶段未引入系统误差时，通过较低程度的预平滑模型即可准确估计 SEE；模型平滑程度（C 参数）为 4 到 6 时，SEE 的估计更为准确。但在这方面也有不同的见解。例如，Puhan, von Davier 和 Gupta（2008）对不可能分数（即 NEAT 中存在个别分数组合的作答数据缺失的情况）研究后发现，剔除不可能分数后，预平滑结果有所改善，但前后差异不大。

以上仅说明，核等值结果受分数分布特征影响，但鉴于核等值采用平滑与连续化处理，在同等条件下，其表现仍会优于其他未考虑离散数据的等值方法。例如，Cid 和 von Davier（2015）通过模拟五种分布形态（对称、正偏、负偏、两种稍微负偏）的测验数据，发现在极端分数处和分数频率较小处，核等值优于 EE。

4 展望

综合来看，未来相关研究可能在如下领域展开：

第一，核等值体系的扩充与完善。

核等值体系以其较强的扩展性，为测验等值方法的创新与发展提供了便利。如前所述，在核等值框架下，von Davier 和 Kong（2005）整合了 Tucker、Levine OSE 与 CE；von Davier 等人（2007）将 Levine 线性观察分数等值与 EE 相结合，提出了混合等值函数；Moses 和 Holland（2008）进一步完善了 EE 方法；Chen 和 Holland（2009）提出了真分数 CEE，并将 Levine TSE 作为其特例；Chen 和 Holland（2010）及 Chen, Livingston 和 Holland（2011）提出了曲线 Levine OSE（Curvilinear Levine OSE），并将传统 Levine 与 Tucker 等值方法作为其特例，整合为观察分数 EE 函数族；von Davier 和 Chen（2013）提出基于锚测验真分数的混合 Levine EE 与 PSE，等等。这些方法的改善与创新，无不受益于核等值的强大扩展能力，故而，研究者有望在其框架下，发现更多准确、高效的等值方法。

第二，核等值流程的更新与发展。

核等值研究流程的不断完善，主要体现在连续化与带宽选择方法两方面：

（1）连续化方法。高斯核被应用于核等值后，众多连续化方法（核函数）也被引入该领域，例如均匀核、Logistic 核、Epanechnikov 核、自适应核，以及基于对数线性模型的方法（Wang, 2007, 2011）。探索将更多优异的随机变量分布函数应用于核等值，不失为有价值的着眼点。

（2）带宽选择方法。Jones, Marron 和 Sheather（1996）从理论角度详细总结了核密度估计的带宽选择方法，然而其后很长时间，核等值领域并未有相关进展。直到 von Davier 等人（2004）系统地提出核等值理论，并采用惩罚方法选择带宽，学者们又开始将目光转移到带宽选择方法。随后出现了重复平滑法、交叉验证法、Silverman 经验准则，以及似然函数方法（Likelihood method; Wallin et al., 2017）等。探索更多的带宽选择方法，以协助研究者在核等值实践中方便、准确地确定密度函数的平滑程度，同样值得关注。

第三，核等值与其他等值方法结合的深入研究。

更深入地探索将核等值与常用及新兴等值方法相结合，可为等值理论发展和实践提供新的视角与选择。

具体来讲，纳入协变量的等值方法无需借助锚题便可匹配考生能力，从而为高利害考试的等值问题提供了解决方案。但从人口统计学信息中选取最有效的协变量，要在保证等值准确性的同时，又使计算简便、快捷，需要更多研究支撑（Wallin & Wiberg, 2016）。同时，Logistic 回归模型与数据拟合程度，及其他模型的选取，对等值结果的影响，也尚待研究。采用核等值方法，对倾向分数分布进行连续化处理，或许可进一步提高等值准确性（Wallin & Wiberg, 2019）。

在 IRT 观察分数核等值中，采用具有较强稳健性的 IRT 模型拟合作答数据（Kolen & Brennan, 2014）；而核等值可通过连续化减小等值误差。那么，当 IRT 模型与数据不匹配且样本量较小时，该方法是否可在一定程度上弥补因 IRT 模型数据不拟合带来的不足？同时，由于模拟运算程序较为耗时，Andersson（2016）仅在 25 道题的测验中探究了 IRT 观察分数核等值情况，故而该方法在较长测验上的表现仍需更多研究验证。

第四，核等值与常用等值方法的比较。

此部分已经在“研究进展”处详述。概括而言，核等值方法的等值准确性与 CTT 及 IRT 等值方法相当，甚至更好，但众多研究结论并不一致。这可能是因为各研究控制的自变量不尽相同，导致其对结果的交互影响不能得到较好控制与准确解释（Andersson, 2016; Andersson & von Davier, 2014; Liang & von Davier, 2014; von Davier & Chen, 2013; Wiberg et al., 2014）。未来应开展更多模拟研究，以操纵可能影响等值结果的变量，排除非研究目的的无关变量对等值结果的干扰，从而探究不同等值方法间的异同。同时，相对于核等值与 CTT 等值方法，其与 IRT 等值方法的比较研究少之又少，未来可采用更为系统、全面的研究设计，重点探究后两者之间的表现差异。

第四，核等值软件的更新。

在核等值发展前期，ETS 先后开发了用于核等值研究的核等值软件（ETS, 2007b）与 GENASYS 软件（ETS, 2007a），但现已不再公开。随着 R 软件（R Core Team, 2017）在社会科学领域的广泛应用，专门用于处理核等值的 kequate 软件包（Andersson et al., 2013）与功能较为综合的 SNSequate 软件包（González, 2014）相继问世。但 kequate 软件包无法处理内锚设计数据，且常因函数算法问题报错； SNSequate 软件包因几乎涵盖常用等值方法，其对核等值方法的支持略显不足。再者，核等值中复杂的矩阵操作和运算，使得研究者完全自编程序开展等值研究面临巨大困难，得不偿失。故而，继续更新或开发更为功能全面、运算高效、操作方便的软件或软件包，可为该领域的研究和实践提供极大便利。

致谢：感谢华南师范大学心理学院博士后（在站）黄菲菲审阅英文摘要并提出建议。感谢心理学院硕士研究生黄丽芳、袁琪婷对公式的编辑。

参考文献

陈俊丽. (2008). *核等值与其它等值方法的比较研究*(硕士学位论文). 北京语言大学.

关丹丹, 景春丽. (2018). 新高考改革背景下不分文理的数学成绩差异研究. *数学教育学报, 27*(4), 31–34.

罗莲. (2008a). *基于 HSK 数据对核等值法与其他等值方法的比较研究*(博士学位论文). 北京语言大学.

罗莲. (2008b). 一种新的等值方法：核等值法. *心理学探新，28*(2), 69–74.

张敏强, 胡晖. (1988). 略论测验等值的理论、方法和应用. *华南师范大学学报(社会科学版)*, (4), 113–118.

Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement, 53*(4), 459–477.

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software, 55*(6), 1–25.

Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement, 51*(3), 223–238.

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika, 82*(1), 48–66.

Arıkan, Ç. A., & Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *International Journal of Assessment Tools in Education, 5*(3), 417–427.

Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating

under the generalized kernel equating framework. *Journal of Educational Measurement, 49*(3), 269–284.

Chen, H., & Holland, P. (2009). Construction of chained true score equipercentile equatings under the kernel equating (KE) framework and their relationship to Levine true score equating. *ETS Research Report Series, 2009*(1), i–15.

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika, 75*(3), 542–557.

Chen, H. H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In S. E. Fienberg, & W. J. van der Linden (Series Eds.) & A. A. von Davier (Vol. Ed.), *Statistics for social and behavioral sciences: Statistical models for test equating, scaling, and linking* (pp. 185–200). New York City, NY: Springer.

Choi, S. I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating standard errors in equipercentile equating* (Unpublished doctorial dissertation). University of Illinois at Urbana-Champaign.

Cid, J. A., & von Davier, A. A. (2015). Examining potential boundary bias effects in kernel smoothing on equating: An introduction for the adaptive and Epanechnikov kernels. *Applied Psychological Measurement, 39*(3), 208–222.

De Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement, 42*(2), 155–168.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97.

Dorans, N. J., & Puhan, G. (2017). Contributions to score linking theory and practice. In B. Veldkamp, & M. von Davier (Series Eds.) & R. E. Bennett, & M. von Davier (Vol. Eds.), *Methodology of educational measurement and assessment: Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 79–132). Cham, Zug, Switzerland: Springer.

Duong, M., & von Davier, A. A. (2008, March). *Kernel equating with observed mixture distributions in a single-group design*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

ETS. (2007a). *GENASYS* [Computer software]. Princeton, NJ: Author.

ETS. (2007b). *KE Software* [Computer software]. Princeton, NJ: Author.

Godfrey, K. E. (2007). *A comparison of kernel equating and IRT true score equating methods* (Unpublished doctorial

dissertation). The University of North Carolina at Greensboro.

González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software, 59*(7), 1–30.

González, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics & Data Analysis, 89*, 222–244.

González, J., & von Davier, A. A. (2016). An illustration of the Epanechnikov and adaptive continuization methods in kernel equating. In L. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. -C. Wang (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 196. Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 253–262). Cham, Zug, Switzerland: Springer.

Grant, M. C., Zhang, L., & Damiano, I. (2009). An evaluation of kernel equating: Parallel equating with classical methods in the SAT subject tests™ program. *ETS Research Report Series, 2009*(1), i–25.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics, 12*(3), 971–988.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*(3), 254–273.

Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement, 51*(2), 201–211.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133–183.

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design. *ETS Research Report Series, 2006*(1), i–38.

Jiang, Y., von Davier, A. A., & Chen, H. (2012). Evaluating equating results: Percent relative error for chained kernel equating. *Journal of Educational Measurement, 49*(1), 39–58.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association, 91*(433), 401–407.

Kim, H. Y. (2014). *A comparison of smoothing methods for the common item nonequivalent groups design* (Unpublished doctorial dissertation). University of Iowa, Iowa City.

Kim, S., & Lu, R. (2018). The pseudo-equivalent groups approach as an alternative to common-item equating. *ETS Research Report Series, 2018*(1), 1–13.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices*. New York City, NY: Springer Science & Business Media.

Lee, Y. H., & von Davier, A. A. (2008). Comparing alternative kernels for the kernel method of test equating: Gaussian, logistic, and uniform kernels. *ETS Research Report Series, 2008*(1), i–26.

Lee, Y. H., & von Davier, A. A. (2011). Equating through alternative kernels. In S. E. Fienberg, & W. J. van der Linden (Series Eds.) & A. A. von Davier (Vol. Ed.), *Statistics for social and behavioral sciences: Statistical models for test equating, scaling, and linking* (pp. 159–273). New York City, NY: Springer.

Leôncio, W., & Wiberg, M. (2017). Evaluating equating transformations from different frameworks. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 233. Quantitative psychology: The 82nd annual meeting of the Psychometric Society, Zurich, Switzerland, 2017* (pp. 101–110). Cham, Zug, Switzerland: Springer.

Liang, T., & von Davier, A. A. (2014). Cross-validation: An alternative bandwidth-selection method in kernel equating. *Applied Psychological Measurement, 38*(4), 281–295.

Liu, J., & Low, A. C. (2007). An exploration of kernel equating using SAT® data: Equating to a similar population and to a distant population. *ETS Research Report Series, 2007*(1), i–22.

Liu, J., & Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT® data. *Journal of Educational Measurement, 45*(4), 309–323.

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics, 40*(3), 227–253.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lu, R., & Guo, H. (2018). A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking. *ETS Research Report Series, 2018*(1), 1–16.

Mao, X. (2006). *An investigation of the accuracy of the estimates of standard errors for the kernel equating functions* (Unpublished doctorial dissertation). University of Iowa, Iowa City.

Meng, Y. (2012). *Comparison of kernel equating and item response theory equating methods* (Unpublished doctorial dissertation). University of Massachusetts Amherst.

Moses, T., & Holland, P. (2007). Kernel and traditional equipercentile equating with degrees of presmoothing. *ETS Research Report Series, 2007*(1), 1–39.

Moses, T., & Holland, P. (2008). Notes on a general framework for observed score equating. *ETS Research Report Series, 2008*(2), i–34.

Moses, T., Yang, W. L., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores.

Journal of Educational Measurement, 44(2), 157–178.

Norman Dvorak, R. L. (2009). *A comparison of kernel equating to the test characteristic curve method* (Unpublished doctorial dissertation). University of Nebraska, Lincoln.

Puhan, G., von Davier, A., & Gupta, S. (2008). Impossible scores resulting in zero frequencies in the anchor test: Impact on smoothing and equating. *ETS Research Report Series, 2008*(1), i–26.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sansivieri, V., & Wiberg, M. (2017). IRT observed-score equating with the nonequivalent groups with covariates design. In L. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. -C. Wang (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 196. Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 275–285). Cham, Zug, Switzerland: Springer.

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica, 77*(4), 329–352.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*(3), 261–285.

Underhill, J. L. (2017). *The robustness of kernel equating as non-normality occurs under the equivalent groups design* (Unpublished doctorial dissertation). University of Florida, Gainesville.

van der Linden, W. J. (2010). On bias in linear observed-score equating. *Measurement: Interdisciplinary Research & Perspective, 8*(1), 21–26.

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement, 50*(3), 249–285.

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement, 34*(8), 620–640.

von Davier, A. A. (2011a). An observed-score equating framework. In P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger (Series. Eds.) & N. J. Dorans, & S. Sinharay (Vol. Eds.), *Lecture notes in statistics: Proceedings: Vol 202. Looking back: proceedings of a conference in honor of Paul W. Holland* (pp. 221–238). New York City, NY: Springer.

von Davier, A. A. (2011b). A statistical perspective on equating test scores. In S. E. Fienberg, & W. J. van der Linden (Series Eds.) & A. A. von Davier (Vol. Ed.), *Statistics for social and behavioral sciences: Statistical models for test equating, scaling, and linking* (pp. 1–17). New York City, NY: Springer.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*(4), 605–623.

von Davier, A. A., & Chen, H. (2013). The kernel levine equipercentile observed-score equating function. *ETS Research Report Series, 2013*(2), i–27.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating. *ETS Research Report Series, 2007*(1), i–19.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data. *ETS Research Report Series, 2006*(1), i–31.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York City, NY: Springer-Verlag.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the nonequivalent groups design. *Journal of Educational and Behavioral Statistics, 30*(3), 313–342.

Wallin, G., Häggström, J., & Wiberg, M. (2017). How to select the bandwidth in kernel equating-An evaluation of five different methods. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 233. Quantitative psychology: The 82nd annual meeting of the Psychometric Society, Zurich, Switzerland, 2017* (pp. 91–100). Cham, Zug, Switzerland: Springer.

Wallin, G., & Wiberg, M. (2016). Nonequivalent groups with covariates design using propensity scores for kernel equating. In L. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. -C. Wang (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 196. Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 309–319). Cham, Zug, Switzerland: Springer.

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics, 44*(4), 1–25.

Wang, T. (2007). *An alternative continuization method to the kernel method in von Davier, Holland and Thayer's (2004) test equating framework* (No. 11). Retrieved Jan 8, 2020, from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/publications-and-data-files

Wang, T. (2011). An alternative continuization method: The continuized log-linear method. In S. E. Fienberg, & W. J. van der Linden (Series Eds.) & A. A. von Davier (Vol. Ed.), *Statistics for Social and Behavioral Sciences: Statistical models for test equating, scaling, and linking* (pp. 141–157). New York City, NY: Springer.

Wang, T., Lee, W. C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632–651.

Wedman, J. (2017). *Theory and validity evidence for a large-scale test for selection to higher education* (Unpublished doctorial dissertation). Umeå University.

Wiberg, M. (2016a). Alternative linear item response theory observed-score equating methods. *Applied Psychological Measurement, 40*(3), 180–199.

Wiberg, M. (2016b). Ensuring test quality over time by monitoring the equating transformations. In L. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. -C. Wang (Vol. Eds), *Springer proceedings in mathematics & statistics: Vol. 196. Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 239–251). Cham, Zug, Switzerland: Springer.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349–361.

Wiberg, M., & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement, 53*(1), 106–125.

Wiberg, M., & van der Linden, W. J. (2011). Local linear observed-score equating. *Journal of Educational Measurement, 48*(3), 229–254.

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement, 51*(1), 57–74.

Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing, 17*(2), 105–126.

Xin, T., & Zhang, J. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement, 39*(1), 44–61.

Kernel equating: A framework of observed score equating

WANG Shaojie; ZHANG Minqiang; LI Tuoyu; LIANG Zhengyan

(School of Psychology, South China Normal University, Guangzhou 510631, China)

Abstract: Kernel equating procedures include pre-smoothing, estimation of the score probabilities, continuization, equating, and evaluation of equating performance. By incorporating linear equating and equipercentile equating methods, kernel equating is more extensible and comprehensive. Pre-smoothing and continuization are distinctive features in kernel equating to reduce the standard error of equating. Standard error of the difference between equating functions are calculated as criterion for evaluating the performances of different kernel equatings. Continuization methods, bandwidth selection methods, etc., can affect the performance of kernel equating. New equating methods based on kernel equating provide an innovative perspective for researchers. Further researchers could focus on extending kernel equating framework by integrating other methods, updating smoothing procedures, and comparative studies.

Key words: kernel equating; continuization; bandwidth selection; new equating methods